

Using full genome sequence to  
compare pathogenic and non-  
pathogenic isolates of *Theileria*  
*orientalis* in Australia

A thesis submitted in fulfilment of the requirements for  
the award of the degree

*Doctor of Philosophy*

*from*

*University of Technology Sydney*

*by*

**Jerald Guojun Yam, BSc (Hons)**

School of life sciences

August 2020

## **Declaration**

### **CERTIFICATE OF ORIGINAL AUTHORSHIP**

I, Jerald G Yam, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Life Sciences at the University of Technology Sydney. This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

Production Note:

**Signature:** Signature removed prior to publication.

**Date:** 11<sup>th</sup> August 2020

## **Acknowledgements**

I wish to express my sincere appreciation to my supervisors, Dr. Cheryl Jenkins, Dr. Daniel Bogema and Prof Steven Djordjevic for their support, patience, wisdom, guidance and the opportunity for completing this Ph.D. project. To Cheryl, thank you for the encouragement; taking the time to review my work and progress; and always providing valuable feedback. To Daniel, thank you for the support throughout my candidature, especially with bioinformatic tasks, this could not have been achieved without your guidance. To Steve, thanks for being my principal supervisor and handling my scholarships and travel grants.

I would like to thank AUSGEM for the scholarship stipend as well as the University of Technology Sydney (UTS) for the international research scholarship.

I am also grateful to Dr. Michael Liu, formally from UTS and the Ramaciotti Centre for Genomics for their Illumina sequencing services.

All of my research was conducted at the Elizabeth Macarthur Agricultural Institute (EMAI), and I would like to acknowledge Dr. Melinda Niedermayer and Shayne Fell for their technical assistance. Also, everyone in the Microbiology and Parasitology section as well as other sections at EMAI for the memorable and joyful experience.

A Ph.D. can be mentally challenging, and I would like to thank my fellow teammates from CYL dragonboat club for their encouragement, support and providing me the much-needed work-life balance.

Words cannot express how grateful I am to my parents, Yvonne and Jonathan Yam. Thank you for the sacrifices and opportunities for me to pursue my dreams. Also, to my sister, Jillian Yam, thank you for the moral support. Last but definitely not least, to my wife, Paulina Sudjarmiko-Yam, thank you for the encouragement, understanding and patience. It was definitely not an easy journey, but your unwavering support has given me the confidence to overcome many obstacles throughout my candidature.

## **Format of thesis**

The Ph.D. thesis presented here for examination is in the format of compilation. This study includes a published introduction chapter and four research chapters in paper publication format. The rationale for this thesis format is to enable the reader to have a better flow of information. Each research chapter contributes and builds on the following chapter to achieve the various aims of this study and it also facilitates the transition to publication.

## **List of publications**

- 1) Yam, J., Bogema, D.R. and Jenkins, C., 2018. Oriental Theileriosis. In: Ticks and Tick-Borne Pathogens. IntechOpen. DOI: 10.5772/intechopen.81198. \*
- 2) Yam, J., Gestier, S., Bryant, B., Campbell-Ward, M., Bogema, D., Jenkins, C., 2018. The identification of *Theileria bicornis* in captive rhinoceros in Australia. Int. J. Parasitol. Parasites Wildl. 7, 85-89.

\* Article reproduced with permission from publisher.

## Table of contents

<b>Declaration</b> .....	ii
<b>Acknowledgements</b> .....	iii
<b>Format of thesis</b> .....	iv
<b>List of publications</b> .....	iv
<b>List of Figures</b> .....	ix
<b>List of Tables</b> .....	x
<b>Abbreviations</b> .....	xii
<b>Abstract</b> .....	xiv
<b>Statement on ethics</b> .....	xv
<b>1. Chapter One: Oriental theileriosis</b> .....	1
<b>Declaration</b> .....	2
<b>1.1 Introduction</b> .....	3
<b>1.2 Taxonomy of <i>T. orientalis</i></b> .....	4
1.2.1 Taxonomic history of <i>T. orientalis</i> .....	4
1.2.2 Taxonomic classification using molecular techniques.....	5
1.2.3 Current taxonomic state of <i>T. orientalis</i> .....	7
<b>1.3. Epidemiology</b> .....	9
1.3.1 Case studies in countries with severe outbreaks: Japan, Australia, New Zealand.....	14
1.3.2 Global distribution of <i>T. orientalis</i> .....	16
1.3.3 Vectors of <i>T. orientalis</i> .....	18
<b>1.4. Lifecycle and transmission</b> .....	20
<b>1.5. Pathogenesis</b> .....	25
<b>1.6. Clinical disease, infection dynamics and the immune response</b> .....	27
<b>1.7. Diagnosis</b> .....	30
<b>1.8. Treatment and control of <i>T. orientalis</i></b> .....	33
1.8.1 Chemotherapy .....	33
1.8.2 Vector control and animal management.....	34
1.8.3 Vaccine development .....	36
<b>1.9 Current state of <i>T. orientalis</i> genomics</b> .....	38
<b>1.10 Application of genomics for <i>T. orientalis</i></b> .....	39
<b>1.11 Rationale and aims of this study</b> .....	41
<b>1.12 Sample data and rationale for selection</b> .....	43

<b>2. Chapter Two: A comparison of methods for the enrichment of <i>T. orientalis</i> DNA from bovine blood samples for next generation sequencing.....</b>	<b>45</b>
<b>2.1 Introduction .....</b>	<b>46</b>
<b>2.2 Materials and Methods.....</b>	<b>47</b>
2.2.1 Blood samples.....	47
2.2.2 Leukocyte depletion .....	48
2.2.3 DNA extraction.....	49
2.2.4 Selective depletion of CpG methylated DNA .....	49
2.2.5 Selective whole genome amplification (SWGA).....	49
2.2.6 Quantitative PCR .....	51
2.2.7 Determining host-to-parasite DNA ratio.....	52
2.2.8 Illumina Sequencing and analysis.....	52
<b>2.3 Results .....</b>	<b>54</b>
2.3.1 Host-to-parasite DNA quantification .....	54
2.3.2 Leukocyte filtration efficiency.....	54
2.3.3 NEBNext microbiome DNA enrichment .....	56
2.3.4 SWGA .....	57
2.3.5 Post-sequencing analysis.....	60
<b>2.4 Discussion.....</b>	<b>62</b>
<b>2.5 Conclusion.....</b>	<b>65</b>
<b>3. Chapter Three: Evaluation of <i>T. orientalis</i> genome assembly methods using nanopore sequencing and analysis of variation between genomes.....</b>	<b>67</b>
<b>3.1 Introduction .....</b>	<b>68</b>
<b>3.2 Materials and Methods.....</b>	<b>69</b>
3.2.1 Illumina sequencing .....	69
3.2.2 Genomic DNA extraction.....	70
3.2.3 Pulsed-field gel electrophoresis (PFGE) .....	71
3.2.4 Nanopore library preparation and sequencing .....	71
3.2.5 Sequence quality assessment .....	72
3.2.6 Genome assembly .....	72
3.2.7 Genome annotation .....	73
3.2.8 Ortholog clustering, phylogeny, gene presence/absence and average nucleotide identity .....	73
<b>3.3 Results.....</b>	<b>74</b>

3.3.1 Comparison of Nanopore reads to Illumina reads .....	74
3.3.2 Draft assemblies of individual assemblers – Nanopore long reads .....	76
3.3.3 Draft assemblies of individual assemblers – hybrid .....	76
3.3.4 Manual genome scaffolding .....	77
3.3.5 Chromosome structures of <i>T. orientalis</i> .....	78
3.3.6 Genome annotation statistics .....	79
3.3.7 Ortholog clustering, phylogeny, gene presence/absence and average nucleotide identity .....	82
<b>3.4 Discussion</b> .....	85
<b>3.5 Conclusion</b> .....	89
<b>4. Chapter Four: A parasite variant pipeline (PVP) to detect and filter high-quality SNPs of <i>T. orientalis</i>.</b> .....	90
4.1 Introduction .....	91
4.2 Materials and Methods.....	93
4.2.1 Variant calling.....	93
4.2.2 Variant filtering.....	96
4.2.3 Validating and testing PVP.....	97
4.3 Results .....	98
4.3.1 Validating and testing PVP with Test-set.....	98
4.3.2 Testing with real-world data.....	98
4.3.3 Additional tools to variant calling in PVP.....	99
4.3.4 Pipeline revalidation.....	101
4.3.5 Comparison of SNPs detected with two different aligners .....	102
4.3.6 Validation metrics of variant calling at varying coverages .....	105
4.3.7 PVP variant filtering .....	106
4.4 Discussion.....	109
4.5 Conclusions .....	112
<b>5. Chapter Five: The analysis of pathogenic <i>T. orientalis</i> isolates within Australia to elucidate its diversity and population structure.</b> .....	113
5.1 Introduction .....	114
5.2 Materials and Methods.....	115
5.2.1 Sample data and parasite DNA extraction.....	116
5.2.2 Selective depletion of CpG methylated DNA .....	116
5.2.3 Evaluating host-to-parasite ratios .....	117

5.2.4 Illumina sequencing and variant calling with PVP .....	119
5.2.5 Downstream analysis of VCF files, heterozygous allele frequency .....	119
5.2.6 Recombination detection .....	120
5.2.7 Population genetics analysis, principal coordinate analysis and phylogeny	120
5.2.8 Deconvolution of multiple haplotype infections .....	122
<b>5.3 Results</b> .....	123
5.3.1 Analysis of variants post variant calling with PVP .....	123
5.3.2 Analysis of variants post quality filtering with PVP.....	123
5.3.3 Recombination analysis.....	125
5.3.4 Population genetics analysis, principal coordinate analysis and phylogeny	127
5.3.5 Deconvolution of multiple haplotype infections .....	130
<b>5.4 Discussion</b> .....	132
<b>5.5 Conclusion</b> .....	136
<b>6. General discussion and concluding statements</b> .....	137
<b>7. References</b> .....	145
<b>8. Appendix I - List of buffer compositions</b> .....	169
<b>9. Appendix II - Supplementary data</b> .....	171
9.1 Chapter Two .....	171
9.2 Chapter Three .....	176
9.3 Chapter Four .....	178
9.4 Chapter Five .....	182



## List of Figures

<b>Figure 1.1:</b> Map of Australia (A) and New Zealand (B) showing the extent of spread of theileriosis during the recent disease incursions in each respective country .....	16
<b>Figure 2.1:</b> The comparison of host-to-parasite DNA ratios between pre and post LF samples demonstrates LF efficiency.....	55
<b>Figure 2.2:</b> The comparison of host-to-parasite DNA ratios between pre- and post-treatment of samples with the NEBNext microbiome kit (n = 3).....	57
<b>Figure 2.3:</b> SWGA efficacy represented by host-to-parasite ratios and the coverage depth distribution of both non-SWGA and SWGA samples.....	58
<b>Figure 2.4:</b> The comparison of coverage depth across reference and coverage histogram.....	59
<b>Figure 3.1:</b> Nucmer alignments of the three <i>T. orientalis</i> strains Robertson, Fish Creek and Goon Nure against the <i>T. orientalis</i> Shintoku reference genome. ....	80
<b>Figure 3.2:</b> Maximum-likelihood tree of Piroplasmida whole genome protein sequences inferred with concordance factors with IQ-TREE 2 using 848 concatenated protein sequences from single copy genes. ....	84
<b>Figure 4.1:</b> Flow chart of PVP.....	95
<b>Figure 4.2:</b> Flow chart showing PVP validation steps and test dataset generated to facilitate validation of PVP using simulated mutated genomes. ....	100
<b>Figure 4.3:</b> Bar graphs of Modified-set SNPs generated by variant calling in PVP and total number of potential SNPs available for variant filtering to achieve high-quality SNPs.....	103
<b>Figure 4.4:</b> Relationship between sensitivity, positive predictive value (PPV) and coverage depth of the simulated mutation datasets of Modified-set. ....	105
<b>Figure 5.1:</b> Sample distribution map and the relevant Illumina sequencing runs on MiSeq and/or NextSeq platforms for each individual sample. .... <b>Error! Bookmark not defined.</b>	
<b>Figure 5.2:</b> Graphs comparing the number of homozygous and heterozygous SNPs called by PVP before variant filtering, between two reference sequences .....	124
<b>Figure 5.3.1:</b> N × N pairwise distance matrix .....	128
<b>Figure 5.3.2:</b> A principal coordinate analysis (PCoA) of 17 isolates .....	129
<b>Figure 5.3.3:</b> An unrooted neighbour-joining tree inferred from genome-wide pairwise distances of 17 isolates.....	130
<b>Supplementary data</b>	
<b>Figure S2.1:</b> Post-sequencing comparison of coverage across reference and coverage histograms between SWGA and non-SWGA samples .....	171
<b>Figure S3.1:</b> QUAST contig Nx plot showing contig lengths across the entire merged genome assembly .....	176

<b>Figure S3.2:</b> Cluster of ortholog groups (COGs) analysis .....	176
<b>Figure S4.1:</b> The analysis of insert sizes of aligners BWA and NGM using TO16-10k-1 as an example .....	178
<b>Figure S5.1:</b> Heterozygous SNP distribution of 17 samples.....	184
<b>Figure S5.2:</b> The outputs of DEploidIBD for each individual sample (n=17).....	186
<b>Figure S5.3:</b> Truncated alignments of concatenated homozygous SNPs of chromosome 1 to 4 to demonstrate recombination .....	196
<b>Figure S5.4:</b> $N \times N$ pairwise distance matrix with calculated genome-wide distances .....	197

## List of Tables

<b>Table 1.1:</b> The global distribution of <i>T. orientalis</i> MPSP genotypes reported in four different host species and the possible transmission vectors. ....	10
<b>Table 1.2:</b> The sample ID, location, year of isolation and the specific studies these samples were used for analysis.....	44
<b>Table 2.1:</b> A summary of enrichment methods and sequencing process samples were subjected to in this three-part pilot study.....	53
<b>Table 2.2:</b> The comparison of host-to-parasite ratios with in-silico and qPCR quantitation. ....	61
<b>Table 3.1:</b> Draft assembly results of the four different assemblers trialled .....	75
<b>Table 3.2:</b> Best manual de novo hybrid assembly for each of the three <i>T. orientalis</i> isolates.....	77
<b>Table 3.3:</b> Genome annotation statistics of the three <i>T. orientalis</i> isolates sequenced in this study and the <i>T. orientalis</i> Shintoku reference sequence. ....	81
<b>Table 4.1:</b> Validation results of mutations datasets in Modified-set. ....	104
<b>Table 4.2:</b> PVP variant filter results at $8 \times$ , $16 \times$ and $32 \times$ coverage depths and the comparison of mutations between two mapping tools: BWA and NGM. ....	108
<b>Table 5.1:</b> Detected recombination events across all 4 chromosomes of <i>T. orientalis</i> Ikeda and the respective recombinant sequences. ....	126
<b>Table 5.2:</b> The number of homozygous and heterozygous SNPs in each of the 17 isolates post quality filtering through PVP against the <i>T. orientalis</i> Shintoku reference genome. ....	131
<b>Table S2.1:</b> Concentration of TOPO-hprt1 and Ikeda-MPSP plasmid standards.....	174
<b>Table S2.2:</b> Oligonucleotide sequences for the SYBR and TaqMan.....	175
<b>Table S2.3:</b> Primer set, toset87, generated from the scoring algorithm of the SWGA program .....	175
<b>Table S3.1:</b> Results of Nanopore sequencing.....	177

<b>Table S4.1:</b> Training results of mutation datasets of Test-set. ....	179
<b>Table S4.2:</b> Comparison of aligners, BWA and NGM with mutation dataset TO16-10k of Modified-set. ....	180
<b>Table S4.3:</b> Average number of SNPs called for each mutation dataset .....	181
<b>Table S5.1.1:</b> Summary of Illumina-sequenced <i>T. orientalis</i> Ikeda isolates (n=17) that were aligned against the Japanese Shintoku sequence. ....	182
<b>Table S5.1.2:</b> Summary of Illumina-sequenced <i>T. orientalis</i> Ikeda isolates (n=17) that were aligned against the Australian Robertson sequence. ....	183
<b>Table S5.2:</b> Number of Ikeda clones identified in each DEploid iteration. ....	195

## Abbreviations

AD	Alternate depth
ANI	Average nucleotide identity
BAM	Binary alignment map
BWA	Burrows-Wheeler Aligner
COGS	Cluster of ortholog groups
COX	Cytochrome oxidase
CTL	Cytotoxic T lymphocytes
EDTA	Ethylenediamine tetraacetic acid
EMAI	Elizabeth Macarthur Agricultural Institute
FN	False negative
FP	False positive
GATK	Genome analysis toolkit
GCN	Gene copy number
gDNA	genomic DNA
GPU	Graphics processing unit
HPCC	High performance computer cluster
HS	High sensitivity
HiS	HiSeq
IBD	Identity by descent
IFAT	Indirect fluorescent antibody test
INDEL	Insertion and deletion
ITS	Internal transcribed spacer
LF	Leukocyte filtration
MAF	Minor allele frequency
MCMC	Markov chain Monte Carlo
MDS	Multidimensional scaling
MGW	Molecular grade water
MOPS	3-Morpholinopropane-1-sulfonic acid
MPSP	Major piroplasm surface protein
MiS	MiSeq

NCBI	National Centre for Biotechnology Information
NGM	NextGenMap
NGS	Next generation sequencing
NSW	New South Wales
ONT	Oxford Nanopore Technologies
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PFGE	Pulsed-field gel electrophoresis
PLAF	Population level allele frequency
PPV	Positive predictive value
PVP	Parasite variant pipeline
QLD	Queensland
qPCR	Quantitative polymerase chain reaction
RBC	Red blood cell
RD	Reference depth
RDP	Recombination detection program
RFLP	Restriction fragment length polymorphism
RLB	Reverse line blot
RT	Room temperature
SAM	Sequence alignment map
SNP	Single nucleotide polymorphism
SWG	Selective whole genome
SWGA	Selective whole genome amplification
TE	Tris-EDTA
TN	True negative
TP	True positive
UNSW	University of New South Wales
UTS	University of Technology Sydney
VCF	Variant calling format
WA	Western Australia
WSAF	Within sample allele frequency

## Abstract

*Theileria orientalis* is an apicomplexan haemoparasite that causes oriental theileriosis in cattle. The parasite is an economic burden globally and there are currently no suitable therapies for the disease in Australia. It is vital to understand the genome of *T. orientalis* to aid the development of efficacious vaccines and/or therapeutics. Next generation sequencing provides an opportunity to comparatively study the *T. orientalis* genome to elucidate its diversity and population structure.

Removal of contaminating host DNA is critical for next-generation sequencing (NGS) of intracellular parasites. This thesis investigated the efficacy of *T. orientalis* DNA enrichment on fresh and frozen clinical blood samples using cellulose leukocyte depletion columns, a commercial microbiome enrichment kit, and selective whole genome amplification (SWGA). All methods were demonstrated to be effective but with unique limitations. qPCR assays were developed to estimate host-to-parasite ratio and identified the commercial microbiome kit to be the most feasible enrichment method for this project.

Long read sequencing of three common *T. orientalis* genotypes was achieved with an Oxford Nanopore MinION flow-cell. Using four distinct assembly methods and manual scaffolding, chromosomal-level assemblies of all three genomes was achieved. Genome annotation and comparative analysis revealed species-level, structural and gene content differences between pathogenic (genotype Ikeda, str. Robertson) and non-pathogenic (genotypes Chitose and Buffeli, str. Fish Creek and str. Goon Nure, respectively) strains similar to that of transforming Theilerias, *T. parva* and *T. annulata*.

To generate high-quality SNPs, a Parasite Variant Pipeline (PVP) was developed with stringent quality filters to eliminate low coverage, erroneous or non-coding SNPs. Validation was performed with 270 simulated sequences over three levels of coverage

breadths and real-world data. To account for errors caused by large indel events, PVP was revalidated and validation metrics revealed no false positives as well as > 88% PVP sensitivity over 16 × coverage.

Finally, a study of diversity and recombination within clinical Australian isolates was conducted with methods and tools investigated in the preceding research chapters. Illumina sequencing was performed on clinical samples that were enriched for parasite DNA. Sequencing data was processed through PVP with long read sequenced and previously published reference sequences to detect high-quality SNPs that were used to identify signatures of meiotic recombination and genetic diversity with genomic distance measures. A total of 15,901 high-quality SNPs were detected by PVP and results indicate *T. orientalis* has low sequence diversity in Australia, which suggests potential for the development of a successful vaccine.

## **Statement on ethics**

During the course of my candidature, there were no procedures performed on any animals. All blood samples used in the studies were collected as part of routine clinical investigations by registered veterinarians. Purified piroplasms of the Robertson, Fish Creek and Goon Nure strains were collected in a previous study (Bogema et al., 2018). That study was carried out in accordance with the Australian Code of Practice for the Care and Use of Animals for Scientific Purposes at the Tick Fever Centre, Wacol, Queensland.